CONFIDENCE LIMITS FOR THE EXPECTED VALUE OF AN ARBITRARY BOUNDED RANDOM VARIABLE WITH A CONTINUOUS DISTRIBUTION FUNCTION

BY

T. W. ANDERSON

TECHNICAL REPORT NO. 1

OCTOBER 1, 1969

PREPARED UNDER THE AUSPICES

OF

OFFICE OF NAVAL RESEARCH CONTRACT #N00014-67-A-0112-0030

DEPARTMENT OF STATISTICS

STANFORD UNIVERSITY

STANFORD, CALIFORNIA

1. Introduction.

Consider a random variable X with a continuous cumulative distribution function $F(x)$ such that $F(a) = 0$ and $F(b) = 1$ for known finite numbers a and b $(a < b)$. The distribution function $F(x)$ is unknown. A sample of size n is drawn from this distribution. Confidence limits for the expected value $\mathcal{E}X$ are to be found that hold for all continuous distribution functions with range $[a, b]$.

2. Confidence limits for the mean.

Let $x^{(1)} < x^{(2)} < \ldots < x^{(n)}$ be the ordered observations in the sample of n from $F(x)$, and let $x^{(0)} = a$ and $x^{(n+1)} = b$. The empirical cumulative distribution function in $[a, b]$ is

$$(1) \qquad F_n(x) = j/n, \qquad x^{(j)} \leq x < x^{(j+1)}, \qquad j = 0, 1, \ldots, n;$$

$F_n(x) = 0$ for $x < a$ and $F_n(x) = 1$ for $x > b$.

Let β and γ be numbers (depending on n) such that the probability of

$$(2) \qquad F_n(x) - \beta \leq F(x) \leq F_n(x) + \gamma, \qquad \text{all } x,$$

is $1 - \alpha$, the desired confidence level. Note that $F_n(x) - \beta \leq 0$ for $x < x^{(r+1)}$, where $r = [n\beta]$, the largest integer in $n\beta$, and $1 \leq F_n(x) + \gamma$ for $x^{(n-s)} \leq x$, where $s = [n\gamma]$. Since $0 \leq F(x) \leq 1$,

the left-hand inequality in (2) is effective only for $x^{(r+1)} \leq x \leq b$ and is replaced by 0 for $a \leq x < x^{(r+1)}$; the right-hand inequality is effective only for $a \leq x < x^{(n-s)}$ and is replaced by 1 for $x^{(n-s)} \leq x \leq b$. The inequalities (2) over these ranges are equivalent to

$$(3) \qquad \frac{j}{n} - \beta \leq F(x^{(j)}) \ , \ j = r+1, \ldots, n, \quad F(x^{(j)}) \leq \frac{j-1}{n} + \gamma, \ j = 1, \ldots, n-s.$$

The distribution satisfying the first part of (3) for given $x^{(r+1)}, \ldots, x^{(n)}$ which has the largest mean is the distribution which has a jump of $(r+1)/n - \beta$ at $x^{(r+1)}$, jumps of $1/n$ at $x^{(j)}$, $j = r+2, \ldots, n$, and a jump of β at b . This fact leads to the inequality

$$(4) \qquad \mathcal{E}X \leq \frac{1}{n}\left[(r+1)x^{(r+1)} + \sum_{j=r+2}^{n} x^{(j)}\right] + \beta\left[b - x^{(r+1)}\right] ;$$

Similarly the distribution satisfying the second part of (3) which has the smallest mean is the distribution which has a jump of γ at a , jumps of $1/n$ at $x^{(j)}$, $j = 1, \ldots, n-s-1$, and a jump of $(s+1)/n - \gamma$ at $x^{(n-s)}$; this fact leads to the inequality

$$(5) \qquad \frac{1}{n}\left[\sum_{j=1}^{n-s-1} x^{(j)} + (s+1)x^{(n-s)}\right] - \gamma\left[x^{(n-s)} - a\right] \leq \mathcal{E}X .$$

The inequalities (4) and (5) hold simultaneously with probability $1 - \alpha$ and these furnish the desired confidence limits. The distributions yielding the upper and lower bounds for $\mathcal{E}X$ are the lower and upper bounds to the distribution of X . If we use integration by parts,

(6) $$\mathcal{E}X = \int_a^b x\,dF(x) = b - \int_a^b F(x)\,dx \; ;$$

the bounds for $\mathcal{E}X$ can be verified by integrating the bounds for $F(x)$.

The inequalities (2) constitute confidence limits for the cumulative distribution function. Values of β and γ for specified values of n and $1 - \alpha$ have been given for $\beta = \gamma$ and for β or γ equal to 1 , making the corresponding inequality vacuous. These are the significance points of the two-sided and one-sided Kolmogorov tests; asymptotic and other approximations are available, as well as tables.

If β is an integer divided by n , namely r/n , the inequality (4) is

(7) $$\mathcal{E}\,X \le \frac{1}{n}\left[\sum_{j=r+1}^{n} x^{(j)} + rb\right] = \bar{x} + \frac{1}{n}\left[rb - \sum_{j=1}^{r} x^{(j)}\right] \; ;$$

the upper confidence limit is the mean of the sample, $\bar{x} = \sum_{j=1}^{n} x^{(j)}/n$, with the r smallest observations replaced by the upper bound. If $\gamma = s/n$, (5) is

(8) $$\frac{1}{n}\left[sa + \sum_{j=1}^{n-s} x^{(j)}\right] = \bar{x} - \frac{1}{n}\left[\sum_{j=n-s+1}^{n} x^{(j)} - sa\right] \le \mathcal{E}X \; ;$$

the lower confidence limit is the mean of the sample with the s largest observations replaced by the lower bound.

This development suggests that it is necessary to assume an upper bound to the range of the random variable in order to obtain an upper

confidence limit for the mean and to assume a lower bound to the range to obtain a lower confidence limit. Indeed, in order that the mean exist conditions on the tails of the distribution are needed, but one cannot verify these conditions with a positive probability on the basis of a finite number of observations. Of course, the need for these bounds limits the applicability of the procedure.

3. Confidence limits for other parameters.

Let $g(x)$ be a monotonically (strictly) increasing function over the interval $[a,b]$. Then the distribution satisfying (2) which has the largest $\mathcal{E}g(X)$ is that which has the largest $\mathcal{E}X$, and correspondingly the distribution satisfying (2) with the smallest $\mathcal{E}g(X)$ is that with the smallest $\mathcal{E}X$. The resulting inequalities are

$$\mathcal{E}g(X) \leq \frac{1}{n}\left[(r+1)g(x^{(r+1)}) + \sum_{j=r+2}^{n} g(x^{(j)})\right] + \beta\left[g(b) - g(x^{(r+1)})\right] , \tag{9}$$

$$\frac{1}{n}\left[\sum_{j=1}^{n-s-1} g(x^{(j)}) + (s+1)g(x^{(n-s)})\right] - \gamma\left[g(x^{(n-s)}) - g(a)\right] \leq \mathcal{E}g(X) . \tag{10}$$

The inequalities (9) and (10) will hold simultaneously for all monotonically (strictly) increasing functions $g(x)$. (The inequality (9), for instance, will hold for $b = \infty$ if $g(b)$ is bounded as $b \to \infty$.)

We can apply (9) and (10) to find confidence limits for the

variance $\sigma^2 = \mathcal{E}X^2 - (\mathcal{E}X)^2$ if $a \geq 0$. Then we have bounds simultaneously on $\mathcal{E}X^2$ and $\mathcal{E}X$, say $L_2 \leq \mathcal{E}X^2 \leq U_2$ and $L_1 \leq \mathcal{E}X \leq U_1$. The latter is equivalent to $L_1^2 \leq (\mathcal{E}X)^2 \leq U_1^2$. Thus

$$L_2 - U_1^2 \leq \sigma^2 \leq U_2 - L_1^2 . \tag{11}$$

4. More general bounds.

The confidence bounds (2) for $F(x)$ can be generalized to

$$F_n(x) - \beta_j \leq F(x) , \quad x^{(j)} \leq x < x^{(j+1)} , \quad j = 0, 1, \ldots , n , \tag{12}$$

$$F(x) \leq F_n(x) + \gamma_j , \quad x^{(j)} \leq x < x^{(j+1)} , \quad j = 0, 1, \ldots , n , \tag{13}$$

for $\beta_j \geq 0$, and $\gamma_j \geq 0$, $j = 0, 1, \ldots , n$. Wald and Wolfowitz (1939) have given expressions for the probability of (12) and (13) holding simultaneously.

If $j/n - \beta_j \leq 0$ or $j/n + \gamma_j \geq 1$, the corresponding inequality (12) or (13) is vacuous. For convenience we shall replace each such value of β_j or γ_j to make $j/n - \beta_j = 0$ or $j/n + \gamma_j = 1$. (In particular $\beta_0 = 0$ and $\gamma_n = 0$.) The sequences of values must satisfy $\beta_j - \beta_{j-1} \leq 1/n$, $j = 1, \ldots , n$, and $\gamma_{j-1} - \gamma_j \leq 1/n$, $j = 1, \ldots , n$, in order to have force in (12) and (13), respectively. The inequalities (12) and (13) are equivalent to

$$(14) \qquad j/n - \beta_j \le F(x^{(j)}) \le (j-1)/n + \gamma_{j-1}, \qquad j = 1, \ldots, n.$$

The cumulative distribution function satisfying (12) and (13) for given $x^{(1)}, \ldots, x^{(n)}$ with the largest mean puts weight $1/n + (\beta_{j-1} - \beta_j)$ at $x^{(j)}$, $j = 1, \ldots, n$, and β_n at b. The distribution satisfying (12) and (13) with the smallest mean puts weight γ_0 at a and $1/n + (\gamma_j - \gamma_{j-1})$ at $x^{(j)}$, $j = 1, \ldots, n$. The resulting inequalities on the expected value are

$$(15) \qquad \mathcal{E}X \le \bar{x} + \sum_{j=1}^{n} (\beta_{j-1} - \beta_j)x^{(j)} + \beta_n b,$$

$$(16) \qquad \bar{x} + \gamma_0 a + \sum_{j=1}^{n} (\gamma_j - \gamma_{j-1})\, x^{(j)} \le \mathcal{E}X,$$

The inequalities (4) and (5) are special cases of (15) and (16).

If $B(y)$ and $C(y)$ are monotonically increasing functions from 0 to 1 in $[0,1]$ and $B(j/n) = j/n - \beta_j$ and $C(j/n) = j/n + \gamma_j$, $j = 0, 1, \ldots, n$, then (12) and (13) can be written

$$(17) \qquad B\left[F_n(x)\right] \le F(x) \le C\left[F_n(x)\right], \qquad \text{all } x.$$

The inequalities (15) and (16) on $\mathcal{E}X$ can be found by integration of (17) by parts as in (6). The form (17) may be helpful in finding limiting probabilities. [See Whittle (1961).]

This approach suggests a number of interesting problems. If one anticipates using these inequalities when the distribution sampled from is a given one, how should one choose $B(y)$ and $C(y)$ to minimize the expected length of the confidence interval? How does the expected length of such a nonparametric interval compare with an interval based on a given parametric form? The intervals presented here are also valid when the distribution sampled from has positive probability at some points but the confidence level will be greater than that stated; modifications of the method for this case are being studied. Further results will be reported in another paper.

References

(1) Wald, A., and Wolfowitz, J. (1939). "Confidence limits for continuous distribution functions," Ann. Math. Statist. 10, 105-118.

(2) Whittle, P. (1961). "Some exact results for one-sided distribution tests of the Kolmogorov-Smirnov type," Ann. Math. Statist. 32, 499-505.